Lab 04: Univariate Descriptive Statistics

**Handout date:** Wednesday, September 23, 2020

**Due date:** Wednesday, September 30, 2020, by midnight into the ***SubmitLab04*** link in eLearning.

*This lab counts 4 % toward your total grade.*

**Objectives:** In this lab   
[a] you will evaluate graphically and numerically mixture distributions with their specific properties is; and  
[b] you split and merge a dataset to evaluate group specific means.

**Format of answer:** Your answers (graphs and verbal description) should be handed in as ***hard-copy*** in ***one*** document. Add a running title into the header of the document with the following information: ***your name***, ***Lab04*** and ***page numbers***. Label each answer properly starting with its task number. Maintain the sequence of questions. Format any code and computer output properly before inserting it into the document with your answer. -code and text output need to be in a ***monospaced*** font (i.e., fixed-pitch font) such as Courier New so proper spacing and alignments are preserved. Excessive, but irrelevant, output will lead to a deduction of your accumulated points.

# Task 1: Descriptive Statistics and Relation to the Mixture Distributions of Random Variables (2.5 points)

Recreate several mixture distributions shown in Table 1 with 1,000 observations using the -script **MixtureDistribExercise.r**. The function **makeMixDist( )** generates the data and the function **mixHist( )**plots the a histograms and two box-plots.

Your ***intellectual challenge*** here is to select a proper set parameters for the two parent normal distributions, i.e., their means and standard deviations, and the mixture proportion of both parent distributions, so that you obtain the distributions shown below.

[a] Find approximately the set of parameters that generate the similar histograms of the mixture distribution. Also insert their associated group-wise box-plots. Select approximate the same landscape plot frame for your box-plots. Each plot here has a width of 3.5’’. (1 point)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mixture ***A*** | Mixture ***B*** | Mixture ***C*** | Mixture ***D*** | Mixture ***E*** |

Table 1: Target Mixture Distributions

Note: Record the parameters of each distribution in the tables at task 1 [b] and 1 [c].

Insert the histograms of the mixture distributions and the group-wise box-plots here:

|  |  |  |
| --- | --- | --- |
| Model | Histogram of Mixture Distribution Model | Group-wise Box-Plot |
| *A* |  |  |
| *B* |  |  |
| *C* | *Your plots go in here* | *Picture width ~ 3.5 inches* |
| *D* |  |  |
| *E* |  |  |

[b] Report your selected parameters of each of the mixture distributions models. (0.5 points)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Mixture Proportion |  |  |  |  |
| *A* |  |  |  |  |  |
| *B* |  |  |  |  |  |
| *C* |  |  |  |  |  |
| *D* |  |  |  |  |  |
| *E* |  |  |  |  |  |

[c] Report in a table the estimated statistics describing the joint distribution of each of the five mixture distribution models. (0.5 points)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Mean | Winsorized Mean at 10% | Standard Deviation | Skewness | Kurtosis | Bimodality Index |
| *A* |  |  |  |  |  |  |
| *B* |  |  |  |  |  |  |
| *C* |  |  |  |  |  |  |
| *D* |  |  |  |  |  |  |
| *E* |  |  |  |  |  |  |

[d] Explain how the skewness, kurtosis and bimodality index relate to the selected parameters that you have selected for the five mixture distribution models? (0.5 points)

Model ***A***:

Model ***B***:

Model ***C***:

Model ***D***:

Model ***E***:

# Task 2: Aggregation – Merger – Weighted Mean (1. 5point)

Open the spatial polygon data-frame **tractShp** in the package **TexMix**. Extract the data-frame **tract <- as.data.frame(tractShp)** and continue working with **tract**.

[a] Generate a professionally labeled histogram and box-plot of the variable **PCTUNIVDEG** (caution: there are missing values) as well as a side-by-side box-plot for **PCTUNIVDEG** broken down by the factor **CITYPERI**. ***Describe*** the distribution of the university degree percentage for the sectors of Dallas County and provide ***meaningful*** summary statistics of the university degree percentages within each sector. (0.3 points)

[b] Aggregate the data-frame by the factor **CITYPERI** into a new data-frame with the aggregated statistics **mean**, and **sd** for the variable **PCTUNIVDEG** as well as number of census tracts in each sector (use the **length()**-function). Show your code and the aggregated data-frame with your calculated statistics. Name the variables properly. (0.4 points)

[c] Compare the regular mean of the university degree percentages based on the census tracts with the ***weighted mean*** based on the aggregated sector means. Use the number of census tracts in each sector as weight. Justify verbally why both means do not differ. (0.4 points)

[d] Merge the aggregated information to your census tract data-frame. Show your code and check that the merger was performed properly by showing the first six records (see the  function **head( )** ). (0.4 points)